Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 102 (2016) 106 – 112

Procedia Computer Science

12th International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS 2016, 29-30 August 2016, Vienna, Austria

Development of an intelligent model to estimate the probability of having metabolic syndrome

Nuriye Sancar [a*], Mehtap Tinazli[b]

[a,*]*Department of Mathematics, Near East University, P.O.Box:99138, Nicosia, North Cyprus, Mersin 10 Turkey*

[b]*Faculty of Medicine, Near East University, P.O.Box:99138, Nicosia, North Cyprus, Mersin 10 Turkey*

Abstract

Logistic regression has now become an essential part of medical data analysis that uses a binary-response model. The model is frequently used by epidemiologists as a model for the probability (interpreted as the risk) that an individual will acquire a disease during a specified time period, during which he or she is exposed to a condition (called a risk factor) known to be or suspected of being associated with the disease. The objective is to establish a model using a minimum number of variables, and is also able to identify the relationship between the dependent variable and independent variable. Additionally, the study will determine the risk factors that can lead to the development of metabolic syndrome (MetSyn) and will establish an intelligent and biologically acceptable model for estimating the probability of having the condition, based on the NCEP ATP III criteria. In this study, binary logistic regression analysis has been employed in order to specify the risk factors that affect metabolic syndrome. Metabolic syndrome (MetSyn) is a common metabolic disorder that is increasingly caused by the pervasiveness of obesity in society and diagnosed according to the National Cholesterol Education Program (NCEP) Adult Treatment Panel III (ATP III) Identification[1]. The data has been obtained from the laboratory test results of 321 adult individuals who had consecutively been treated by the Near East University Internal Medicine Department. For this intelligent model, binary logistic regression analysis has been used. The sensitivity, specificity and accuracy rates have been detected as 94.7%, 96.0% and 95.5%, respectively. As a result, homeostatic model assessment (HOMA-IR), uric acid, body mass index (BMI), low-density lipoprotein (LDL) cholesterol, age, smoking, education level (EL) are defined as metabolic syndrome risk factors, the model has been estimated by using those variables in the acquired intelligent model. As a consequence of the research, it has been determined that the key elements that can have an impact are the changeable risk factors, meaning that the illness could be destroyed before it actually occurs, and lifestyle change, that can also prevent the illness.

Peer-review under responsibility of the Organizing Committee of ICAFS 2016

Keywords: Metabolic syndrome; risk factors; logistic regression; odds ratio.

*Corresponding Author:
Nuriye Sancar. E-mail address: nuriye.sancar@neu.edu.tr

Peer-review under responsibility of the Organizing Committee of ICAFS 2016
doi:10.1016/j.procs.2016.09.376

1. Introduction

Regression analysis is a widely used process for obtaining a predictor function for estimating the value of a dependent variable using the independent or predictor variables. In some cases, the outcome (or dependent) variable can be discrete, with two or more possible values. Logistic regression has now become an essential part of medical data analysis that uses a binary-response model. This method of statistical analysis is an extension of multiple regression methods and can be applied where the dependent variable is categorical, meaning that the values of the variable can be assigned to a countable number of categories. In the medical field, a particular outcome could be caused by the presence or absence of a specific disease. In such cases, logistic regression has become an increasingly common method used to estimate the probability that the outcome will occur as a linear function of one or more continuous and/or dichotomous independent variables[2]. The objective is to establish a model using a minimum number of variables, which is acceptable biologically and is also able to identify the relationship between the dependent variable and independent variable. The model is frequently used by epidemiologists as a model for the probability (interpreted as the risk) that an individual will acquire a disease during a specified time period, during which he or she is exposed to a condition (called a risk factor) known to be or suspected of being associated with the disease[3].

Metabolic syndrome (MetSyn) is a common metabolic disorder that is increasingly caused by the pervasiveness of obesity in society[4]. Patients with metabolic syndrome are increasingly likely to develop diabetes mellitus and cardiovascular disease and the risk of mortality from cardiovascular diseases is greater[4]. Metabolic syndrome is diagnosed according to the National Cholesterol Education Program (NCEP) Adult Treatment Panel III (ATP III) Identification[1]. According to the this definition, metabolic syndrome is present if three or more of the following five criteria are met: waist circumference over 102 cm (men) or 88 cm (women), blood pressure over 130/85 mmHg, fasting triglyceride (TG) level over 150 mg/dl, fasting high-density lipoprotein (HDL) cholesterol level less than 40 mg/dl (men) or 50 mg/dl (women) and fasting blood sugar over 100 mg/dl. The aim of this study is to establish an intelligent and biologically acceptable model for estimating the probability of having metabolic syndrome by finding risk factors that can lead to metabolic syndrome based on those criteria. For this intelligent model, binary logistic regression analysis has been applied.

The formula for a logistic regression model[5] is given by

$$\pi(x_i) = P\,(y_i=1/x_i) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_{k-1}x_{k-1}}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_{k-1}x_{k-1}}} \quad (1)$$

where $y_i = \begin{cases} 1, & Having\ MetSyn \\ 0, & Not\ having\ MetSyn \end{cases}$ i=1,2,…,n

e is the base of the natural logarithm, $x_1, x_2, \ldots, x_{k-1}$ are the independent variables (or risk factors), β_0 is the constant term, β_1, β_2, …, β_{k-1} are the coefficients of the independent variables and $P\,(y_i=1/x_i)$ or $\pi(x_i)$ is the probability that the i^{th} individual will have MetSyn.

As is the case with any regression model, valuable information about the relationship between the independent variables to the binary dependent variable can be provided by the regression coefficients β_j in the logistic model by using equation (1). Logistic regression quantifies the relationship between the dichotomous (or binary) dependent variable and the predictors using odds ratios. Since the logit model provides an estimate of the odds ratio, the binary logistic regression is discussed under the logit link function in this study[6]. To describe the concept of odds ratio, odds should be defined as the probability that an event will occur divided by the probability that the event will not happen[3]. An odds ratio is a ratio of two odds and a measure of how much greater (or less) the odds are for subjects possessing the risk factor to experience a particular outcome. Therefore, we can define the odds with the following ratio:

$$\text{odds} = \frac{\pi(x)}{1-\pi(x)} \quad (2)$$

Rearranging by using equation (1), we get

$$\text{odds} = \frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_{k-1}x_{k-1}} \quad (3)$$

$\text{logit}(\pi(x))$ is the natural logarithm of the odds of outcome, so

$$\text{logit}(\pi(x)) = \ln(odds) = \beta_0+\beta_1 x_1 +\beta_2 x_2+\ldots+\beta_{k-1} x_{k-1} \quad (4)$$

The natural log transformation of the odds in the equation (4) is necessary to construct the relationship between a binary outcome variable and its predictors linear

For each unit increase in x_i the odds increase multiplicably by e^{β_j} or odds ratio (OR). An OR>1 denotes a *higher* risk in the exposed group, so the exposure of the factor increases the risk. An OR<1 denotes a lower risk in the exposed group, so the exposure of the factor reduces the risk[7].

2. Materials and methods

This study has been made with the aim of establishing an intelligent and biologically acceptable model for estimating the probability of having metabolic syndrome by determining the risk factors that can lead to people having metabolic syndrome, based on the NCEP ATP III definition. The data has been obtained from the laboratory test results of 321 adult individuals who had been consecutively treated by the Near East University Internal Medicine Department. Waist circumference, blood pressure, body mass index (BMI), fasting glucose, Total cholesterol level (TCL), High density lipoprotein (HDL)-cholesterol, Low density lipoprotein (LDL) -cholesterol, triglycerides, uric acid, hs-CRP (high sensitivity C-reactive protein), smoking, education level (EL), age and homeostatic model assessment–insulin resistance (HOMA-IR), were evaluated for all adults. The individual features (such as smoking condition) were obtained by directly asking the individuals themselves. The number of patients with metabolic syndrome out of the 321 who had been tested by the Department was determined according to the NCEP ATP III Identification[1].

All analyses were conducted using the statistical package SPSS version 20.0 (SPSS for Windows, Chicago, IL, USA). Binary Logistic regression analysis with logit link function was performed to determine the risk factors by Backward Likelihood Ratio LR elimination method. The Backward LR Elimination method is a stepwise selection method that begins with a full model and predictor variables are excluded using the probability of the likelihood ratio statistic based on the maximum partial likelihood estimates[8]. It was also verified whether multicollinearity and outlier problems were present in the data. The statistical test for goodness of fit test was performed using the Hosmer-Lemeshow test. In addition, sensitivity, specificity and accuracy rates were calculated. A p value < 0.05 was considered statistically significant.

Variables included in the study: The dependent variable was the metabolic syndrome status of an individual (*presence* (coded *1*) or *absence* (coded *0*)), denoted by MetSyn. Independent variables were homoeostasis model assessment (HOMA-IR), uric acid, body mass index (BMI), low-density lipoprotein (LDL)-cholesterol, total cholesterol level (TCL), age, smoking, education level (EL), hs-CRP and gender.

Age and education level were evaluated as continuous variables. The rest of the variables were evaluated as categorical variables. Categorical variables were classified as the following:

LDL<130 mg/dL (coded 0), LDL≥130 mg/dL (coded 1) for non-diabetic and non-coronary heart disease patients;
LDL<100 mg/dL (coded 0), LDL≥100 mg/dL (coded 1) for the patients with diabetic and coronary heart disease;
HOMA-IR<2.5(coded 0), HOMA-IR≥2.5(coded 1);Total cholesterol level(TCL)<200 mg/dL(coded 0), TCL ≥ 200 mg/dL (coded 1); uric acid < 7mg/dL (coded 0) for male, < 6mg/dL(coded 0) for female; non-smoking (coded 0), smoking (coded 1); hs-CRP<0.5 mg/dL(coded 0), hs-CRP ≥ 0.5 mg/dL (coded 1);
BMI< 30 kg/m^2 (coded 0), BMI ≥ 30kg/m^2 (coded 1 as obesity); female (coded 0), male (coded 1).

Using the Pearson Chi-square (χ^2) and the likelihood ratio (G^2) tests, each of the independent variables was tested individually for association with the dependent variable. The results can be found in Table 1.

3. Results

120 (37%) patients who were evaluated in the research were found to have metabolic syndrome. 50.3% (163) of the participants were male, 49.7% (161) of the participants were female. 70% of individuals who have metabolic syndrome were male, 30% were female. Possible predictor variables that can be related with the dependent variable MetSyn, which are shown in Table 1. The predictor variables: BMI, LDL, uric acid, HOMA-IR, hs-CRP, age, smoking, education level, were included for the multivariate model as those variables are correlated with the

dependent variable. The results related to the multivariate binary logistic regression model that was established with predictor variables are shown in Table 2.

Table 1. Tests of association between MetSyn and each of the independent variables

	Pearson Chi-Square Test		Likelihood Ratio Test	
Variable	Value	p-value	Value	p-value
Age (age)	181.97	0.000*	228.961	0.000*
Gender (gender)	29.561	0.000*	30.198	0.000*
High sensitivity Creactiveprotein (hs-CRP)	88.146	0.000*	94.722	0.000*
Total Cholesterol Level (TCL)	2.122	0.145	2.119	0.146
Body Mass Index (BMI)	228.998	0.000*	256.990	0.000*
Uric Acid (uricacid)	105.884	0.000*	107.538	0.000*
Homoeostasis model assessment (HOMA-IR)	174.770	0.000*	187.288	0.000*
low-density lipoprotein cholesterol (LDL)	7.675	0.006*	7.468	0.006*
Education Level (EL)	42.096	0.000*	44.021	0.000*
Smoking (smoking)	122.788	0.000*	126.050	0.000*

*significant (p-value<0.05)

Table 2. Binary Logistic regression analysis of Metabolic Syndrome (MetSyn) using Backward Likelihood Ratio Elimination Method

	Covariates	β	Standard Error	Wald statistics	df	p-value	Exp(β)
Step2**	age	0.082	0.021	14.887	1	0.000*	1.086
	gender (female)	1.453	0.688	4.456	1	0.035*	4.277
	HOMA-IR	2.515	0.667	14.228	1	0.000*	12.363
	BMI	3.100	0.694	19.960	1	0.000*	22.198
	Smoking	1.867	0.615	9.223	1	0.002*	6.472
	EL	-0.168	0.084	3.950	1	0.044*	0.845
	LDL	1.300	0.626	4.311	1	0.038*	3.669
	uricacid	1.248	0.635	4.092	1	0.043*	3.483
	constant	-7.481	1.662	20.266	1	0.000*	0.001

β: logistic coefficient of each covariate in the model Exp(β): Exponentiated logistic coefficients or Odds Ratio df: degrees of freedom of Wald statistics

*If the p value is smaller than 0.05, that variable (or variables category) is determined to be a significant risk factor.

**BMI, LDL, uric acid, HOMA-IR, age, smoking, education level, hs-CRP, gender independent variables entered on step 1.

Interpretation of the odds ratios according to the Binary Multivariate Logistic Regression results are as below:

It was concluded that the odds rate for the variable age greater than 1 indicates that the possibility of having metabolic syndrome rate increased with age by 1.086 times. Furthermore, there was an inverse relation between metabolic syndrome and education level, which is a significant risk factor. Therefore, it was determined that the education level reduces MetSyn or in other words, education has a protective effect (OR:0.845< 1). Moreover, the metabolic syndrome rate for an individual who has a high HOMA-IR value is 12.363 times more than those individuals who had a lower HOMA-IR value. Smokers have 6.472 times more risk to develop MetSyn than those who are not smoking. It was concluded that for the individuals who have obesity, metabolic syndrome disease rates are 22.198 times more than who do not have an obesity problem. Furthermore, individuals with a high LDL have

3.669 times more risk of having metabolic syndrome than individuals with a standard level. Individuals who have a high uric acid level are 3.483 times more at risk of developing metabolic syndrome than those with a standard uric acid level. Furthermore, the logistic regression result has demonstrated that the gender variable (male) is an important risk factor with a positive effect on having metabolic syndrome. This result shows that males are 4.277 times more at risk of having metabolic syndrome than females.

The fitted logit model is:

$\text{logit}(\pi) = -7.481 + (3.1 \times \text{BMI}) + (0.082 \times \text{age}) + (1.867 \times \text{smoking}) + (2.515 \times \text{HOMA-IR}) + (1.248 \times \text{uric acid}) + (1.3 \times \text{LDL}) + (1.453 \times \text{gender}) - (0.168 \times \text{EL})$, where π is the estimated probability of metabolic syndrome.

The Hosmer-Lemeshow statistics was compared by chi-square distribution with a significance level of $\alpha = 0.05$ and 8 df. Since $13.342 < \chi^2_{0.05,8} = 15.507$, it was concluded that the model fit was quite good. Also, the Nagelkerke R^2, which explains how much variance of the data is "explained" by the model, was calculated as 0.883. This shows that the independent variables in the model explain having the metabolic syndrome condition in a good rate. The original logistic regression equation can be transformed to show the estimated probability of having metabolic syndrome by the equation (1).

For example, we can calculate a the probability of metabolic syndrome occurrence of a 40 year old male , with 8 years education, who exceeds the obesity limit (BMI>30kg/m^2), is a smoker, exceeds the determined HOMA-IR value and who has standard uric acid and LDL values (Table 3).

Table 3. Probability estimate with Logistic regression model

Characteristics of an individual	The values of coefficients β
Gender (male), coded 1	1.453 x 1= 1.453
Age (40)	0.082 x 40= 3.28
Education Level (8 years)	-0.168 x 8= -1.344
HOMA-IR (7.1), coded 1	2.515 x 1= 2.515
BMI (35), coded 1	3.1 x 1= 3.1
Uric acid (4), coded 0	1.248 x 0= 0
LDL (80), coded 0	1.3 x 0= 0
Smoking, coded 1	1.867 x 1= 1.867
Constant	-7.481 Total:3.39

$$\pi = P\,(y=1/x) = \frac{e^{3.39}}{1+e^{3.39}} = 0.967 \quad (5)$$

Based on the equation (5), the probability of having metabolic syndrome was estimated as 96.7%, in male individuals who match the specifications listed above. From the equation (6), the probability of having metabolic syndrome was estimated as 87.4%, in female individuals who have the specifications mentioned above.

$$\pi = P\,(y=1/x) = \frac{e^{1.937}}{1+e^{1.937}} = 0.874 \quad (6)$$

This result shows that a male individual's probability of having metabolic syndrome with those parameters is higher than in comparison with a female individual's probability of having metabolic syndrome.

4. Discussion

Our aim in this study was to find risk factors that cause metabolic syndrome and to establish an intelligent and biologically acceptable model for estimating the probability of having metabolic syndrome, based on the NCEP ATP III criteria. Metabolic syndrome is defined by a cluster of risk factors that are associated with diabetes and cardiovascular disease[4]. In our study, significant parameters that have effects on MetSyn has been revealed and a logistic regression model has been established.

The result of the analysis has shown that the risk of having metabolic syndrome in males was higher than in females. The same rate of female and male adults have been included in the study and consequently, this study concluded that the smoking and obesity rates of males were statistically significantly higher than for females ($p<0.001$). Meanwhile, taking into account the individuals with high rates of HOMA-IR and uric acid values included in the model, males were statistically significantly more likely thanfemales to have metabolic syndrome ($p<0.001$). However, contrary to these results, studies in Turkey have concluded that having metabolic syndrome is a more frequent problem for females[9].

In the results, aging increases the risk of having MetSyn and a higher education level decreases the MetSyn risk. It has been shown in many studies that age is an important factor in the generation of cardiometabolic risk factors[10,11]. Identifying the low education condition in this study as a risk factor that increases MetSyn indicates the importance of developing the education level to promote healthy lifestyles, which is important factor in preventing MetSyn occurrence.

Smoking has negative effects on health, particularly on MetSyn, cardiovascular diseases and cancer[12]. In our study, the effect of smoking on increasing the risk of having MetSyn, has been determined to be in line with the literature[12]. Furthermore, 19.3% of the female participants and 44.2% of the male participants were smokers. Consequently, in the case of smoking, which is a changeable risk factor,it is recommended that the authorities should prepare and implement programs for smoking prevention.

In order of other importance, the risk factors (HOMA-IR, uric acid, LDL) are variables that could also be accepted in our model as biological. This is because some research studies have indicated that there is a relation between those mentioned factors and having metabolic syndrome[15,16].

As a result of the analysis, it is concluded that the most important risk factor is obesity. This result harmonizes with the literature because the risk of having metabolic syndrome is closely linked to obesity[4]. Obesity prevalence shows a global increase in parallel with changing lifestyles [13,14]. Apart from changing people's lifestyles, there is no other agent that will provide effective treatment for metabolic syndrome. Therefore, the most appropriate treatments to provide are weight loss programs, regular exercise, healthy dieting and quitting smoking. Furthermore, the ever-increasing rate of metabolic syndrome is explained by inactivity and consuming more energy more than is used, which are the main features of the urban lifestyle[14].

5. Conclusion

In this study is, an intelligent and biologically acceptable model has been established for estimating the probability of having metabolic syndrome, based on the NCEP ATP III criteria. As a result, the BMI, LDL, uric acid, HOMA-IR, age, smoking, education level, variables have been identified as risk factors that affect the probability of having MetSyn. Obesity and smoking are changeable risk factors have been found to be the most important risk factors. The major precaution to be taken on this subject is to raise the awareness of the public. The individuals who have obesity and (or) are smokers, should know that they are under greater risk in terms of hypertension, coronary artery disease and type 2 diabetics. This is because metabolic syndrome is a condition that increases risk of heart diseases and diabetics due to the aging process[4]. The risk factors which causes this disease, should be eliminated before it has the opportunity to develop. Changing people's lifestyles, which can prevent the occurrence of diseases, has vital importance.

References

1. Third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults -Adult Treatment Panel III. Final report. *Circulation* 2002;**106**:3143–3421.
2. Hosmer Jr DW, Lemeshow S and Sturdivant RX. *Applied logistic regression*. 3rd ed. John Wiley & Sons; 2013.

3. Daniel WW. *Biostatistics: a foundation for analysis in the health sciences*. 8th ed. Wiley; 2005.
4. Grundy SM, Brewer HB, Cleeman JI, Smith SC, Lenfant C. Definition of metabolic syndrome report of the National Heart, Lung, and Blood Institute/American Heart Association Conference on scientific issues related to definition. *Circulation* 2004; **109**: 433-438.
5. Harrell FE Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer; 2001.
6. Kudakwashe M and Yesuf KM. Application of Binary Logistic Regression in Assessing Risk Factors Affecting the Prevalence of Toxoplasmosis. *American Journal of Applied Mathematics and Statistics* 2014; **2**:357-363.
7. Scotia N. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry* 2010; **19**: 227.
8. Cokluk O. Logistic Regression: Concept and Application. *Educational Sciences: Theory and Practice* 2010; **10**: 1397-1407.
9. Onat A, Uyarel H, Karabulut AS, Doğan Y, Can G. Halkımızda abdominal obezitede risk faktörü kümelemeleri ve demografik dağılım. *Türk Kardiyol Dern Arş* 2005;**33**:195-203.
10. Onat A, Sansoy V. Halkımızda koroner hastalığın başsuclusumetabolik sendrom: Sıklığı, unsurları, koroner risk ile ilişkisi ve yuksek risk bileşenleri. *Turk Kardiyol Dern Arşivi* 2002; **30**: 8-15.
11. Cameron AJ, Shaw JE, Zimmet PZ. The metabolic syndrome: prevalence in worldwide populations. *Endocrinol Metab Clin N Am* 2004; **33**: 351-75.
12. Miyatake N, Wada J, Kawasaki Y, Nishii K, Makino H, Numata T. Relationship between metabolic syndrome and cigarette smoking in the Japanese population. *Intern Med* 2006;**45**:1039-43
13. Hwang LC, Bai CH, Chen, CJ. Prevalence of obesity and metabolic syndrome in Taiwan. *Journal of the Formosan Medical Association* 2006;**105**: 626-635.
14. Li TY, Rana JS, Manson JE, Willett WC, Stampfer MF, Colditz GA, et al. Obesity as compared with physical activity in predicting risk of coronary heart disease in women. *Circulation* 2006; **113**: 499-506.
15. Jeppesen J, Hansen TW, Rasmussen S, Ibsen H, Torp-Pedersen C, Madsbad S. Insulin resistance, the metabolic syndrome, and risk of incident cardiovascular disease: a population-based study. *Journal of the American College of Cardiology* 2007; *49*: 2112-2119.
16. Silva HAD, Carraro JCC, Bressan J, Hermsdorff HHM (2015). Relation between uric acid and metabolic syndrome in subjects with cardiometabolic risk. *Einstein (Sao Paulo)* 2015; **13**: 202-208.